ONO5424

Return to OISE

PROJECT REPORT

Progress Report: _____ Final Report: XXX Project Number: 85-1050

Small-Scale: _____ Large-Scale: XXX Funding Year: 1987-1988

Transfer Grant: XXX (Priority: TG4) OISE: _____

New: XXX Continuing: _____ Terminated: XXX Re-funded: _____

Project Title: The Analysis and Display of Educational Assessment Survey Data

Investigator(s): Richard Wolfe

Budget Unit: FSR(Educ.Eval.Centre) Date Rec'd in F&SR: June 88

ABSTRACT

For large-scale projects, please provide in the space below an abstract of not more than 300 words, single-spaced, written in a language that may be understood by the informed layperson. For small-scale projects, an abstract per se may not be appropriate, and the report itself may be started below.

In order to study the analysis and display of educational assessment survey data, an advanced computing facility first had to be built. This is composed of a set of linked computing and programming systems that facilitate experimental work in statistical data analysis. They are interactive, in that one can instantly view data from different perspectives, alter parameters and see their effects on the fit of statistical models, zoom in on unusual points, or add and subtract "noise" to data to observe the effects of error and misspecification. The principal educational research data explored in this project were (a) the databank from the Second International Mathematics Study, (b) standardized test data from board-wide testing programs, and (c) data from the school board replications of the IEA studies. The methodological focus was on explanatory models—that is, describing and explaining change in achievement as it occurs within the timeframe and hierarchy of schooling. Some work looked at "microanalysis". How can we explain the gains and losses in students' achievement on a single test item? Other work looked at "macroanalysis". How are the different amounts of achievement observed in different settings (schools, classrooms) related to the characteristics of the students and the characteristics of the settings. Major finding were (1) that our classification of achievement test items is very shaky, and that we need better theoretically-derived linkages between item categories if we are to connect item assessment data with prospects for teaching and learning; (2) schools and classroom are seen to vary in three dimensions: in what kind of student achievement they receive, in what kind of student achievement they produce, above and beyond what they receive, and in how much of an accelerating or flattening effect they apply to the student achievement distribution (in Ontario, the French schools and the private schools are identified as special cases); (3) with respect to standardized tests, there is a small but critical proportion of bad data that must be more carefully detected and treated.

This project was funded by the Ministry of Education under its Block Transfer Grant to OISE.

92-07719/

Final Report

The Analysis and Display of Educational Assessment Survey Data

Richard Wolfe

1. Description of activities

1.1. Development of advanced data analysis and graphics laboratory facility

This project was unusual in that it had budget for equipment and assignment of faculty time, but nothing else. In order to study the analysis and display of educational assessment survey data, an advanced facility first had to be built. And the principal investigator was an expert computer analyst.

Statistical data processing usually takes place in one of three modes.

First, when general-purpose software is not available, or when long-term production is envisioned, there can be special programs written for specific analysis. The efficiency of this approach is questionable, since considerable programming effort has to be devoted to work that will consume a tiny amount of computer time. A related kind of waste of time is to attempt data analysis in an irrelevant computer environment, such as a word processing package or a spreadsheet. These approaches are seen too often in educational settings.

A second and more sensible approach is to use one of the widely available statistical systems, such as SPSS or SAS. These are "packages" in that they provide complete statistical solutions to conventional analysis strategies, with considerable flexibility in defining inputs, performing data transformations, varying statistical parameters, and formatting outputs. The packages are, in essence, programming languages devoted to statistical operations, and they are languages that non-computer-science researchers can master. Versions now exist and are available at reasonable cost for microcomputers as well as mainframes. They can be linked to colour monitors, printers, and plotters for production of conventional statistical graphics, including bar charts, scatterplots, etc.

The third mode for statistical computing involves the use of (a) advanced computing equipment and (b) programming languages. The main problem with conventional statistical packages is that they do not facilitate experimental work in statistical data analysis. They are not very interactive, in that one cannot instantly view data from different perspectives, alter parameters and see their effects on the fit of statistical models, zoom in on unusual points, or add and subtract "noise" to data to observe the effects of error and misspecification. These are essential aspects of exploring new approaches to data analysis.

Since this project required that kind of experimental work, a main activity has been to assembly the hardware and software. The state of this art is, of course, extremely volatile. New hardware capabilities and software systems are announced every day. There are also competing computer platforms relevant to statistical computing—development has proceeded in a rather independent way for mainframes and minicomputers (e.g., IBM and

VAX), for the MS-DOS microcomputer family (e.g., the IBM PC), for the Apple Macintosh computers, and for the UNIX advanced workstations. In fact, to carry out extensive, state-of-the-art exploration of statistical computing methods, one cannot ignore any of these avenues.

The computing facility that has been assembled for the project includes (a) a Macintosh II running the Apple operating system and UNIX., (b) a SUN 3/50 workstation running UNIX, (c) an IBM/AT running MS-DOS, (d) a QMS laser printer running PostScript and HP-emulation, and (e) Ethernet links to the OISE VAX-VMS. From a hardware perspective, this provides an enormous amount of number-crunching capacity, memory for programming, disk storage for data, high-resolution colour screens and large-format high-resolution monochrome screens for display, and high-resolution, high-speed printing and plotting.

The software arrangements for the project include the conventional statistical packages on the VAX, with versions running under UNIX and on the IBM-AT and the Macintosh. More importantly from a research purpose, there is the "S" language on the SUN. This is a main tool for statistical research. It comes from Bell Laboratories and provides a unique combination of statistical operations, graphic commands, and macro programming. The SUN is also an ideal environment for original program development. On the Macintosh, the MacSpin program, which projects three-dimensional data configurations in motion, has proved to be essential for multivariate data exploration. The Macintosh environment is also an excellent place to pull together graphs, text, and diagrams, and to produce reports. The IBM-AT has been used mainly for communication, data transfer, and for a few specialized statistical estimation programs.

This project has been greatly aided by having multiple computing environments. For each aspect of the computing and analysis, the most appropriate computing tools could be used. This saves researcher time. (The speed of statistical computing should be measured by when the analyst starts and stops, not by how long the computer takes.) The linkage of the the different computers has been important. Initially, this was done through regular communications lines and the Kermit program. But now, all the computers are hooked up over the ultra-high-speed Ethernet and share file structures.

1.2. Experimentation with statistical and graphics software

The educational research data explored in this project included (a) the databank from the Second International Mathematics Study, (b) standardized test data from board-wide testing programs, (c) opinion surveys administered by school boards to staff, students, and parents; (d) the databank of the IEA Classroom Environment Study; and (e) data from the school board replications of the IEA studies.

In some cases, the analysis and graphics were limited to conventional approaches, and so they will not be discussed further, except to note that the availability of high-quality statistical graphics should change how we process and present all kinds of educational survey data. There is no excuse now for dull reports and long, badly printed tables of statistics.

As planned, the focus was on cognitive test data—responses to achievement test items—and on multilevel datasets—in which information is available on student achievement and background and on classroom, teacher, and school characteristics. An unanticipated

supplementary focus, toward the end of the project, has been on using open-ended responses.

The steps during experimentation were (a) exploratory analysis and display, especially using the high-speed interactive capabilities of the SUN and the Macintosh; (b) using the "S" language to make quick prototypes of analysis and graphical sequences; (c) using "S", regular programming languages, and packaged programs to carry out statistical simulations and tests; and (d) putting together production versions of procedures. Because the final productions were aimed at research papers and reports, the integration of text, formulas, and graphics was important.

1.3. Derivation of statistical and data analytic methods

While the content was mainly educational achievement, the methodological focus was on explanatory models—that is, describing and explaining change in achievement as it occurs within the timeframe and hierarchy of schooling.

Some work looked at "micro-analysis". How can we explain the gains and losses in students' achievement on a single test item? With the longitudinal data obtained in the Second IEA Mathematics Study, we can actually crosstabulate student knowledge on a mathematics item at the beginning of a school year by knowledge at the end of the year by reports of opportunity to learn the content of the item before and during the school year. Statistical models have been derived for accounting for growth and for dealing with previously unexplored item characteristics, such as the sensitivity of an item to instruction. Graphical procedures have been developed for displaying item statistics in ways that reveal similarities and differences across content and item type domains and that expose unusual items.

Other work looked at "macro-analysis". How are the different amounts of achievement observed in different settings (schools, classrooms) related to the characteristics of the students and the characteristics of the settings. A focus in this project has been on describing and explaining variation in school effects. That is, schools and classrooms (teachers) differ in how much achievement is produced. But they also vary in their characteristic pattern of transforming inputs into outputs. For example, a school that places great emphasis on remediation will tend to flatten the regression of outputs on inputs, while a school that emphasizes enrichment will do the opposite. In this project, techniques were developed for looking at how schools and classrooms are distributed in the three dimensions of input, output, and regression. In particular, the school-classroom effects problem was approached statistically through robust regression and graphically through three-dimensional plotting. Simulation and "jack-knife" analysis was used to assess statistical stability of the estimation.

1.4. Analysis of open-ended response data

As mentioned above, there has been a supplementary focus on the analysis of open-ended responses.

For purposes of economy of scoring, at least, there has been a practically total focus in the educational achievement business on multiple-choice questions. Similarly in opinion surveys, most data are collected with using questions with fixed options. Usually the

open-ended information that is collected is reported with a very slight sample or it is just thrown away.

The opportunity arose during the project to use the advanced equipment to develop coding procedures for a particular set of open-ended response data. Specifically, a school board had collected opinion data from staff and students and there were hundreds of different responses to each question. The problem was to code the responses so that a statistical presentation could be made. The procedure developed was to type the exact responses into the computer, and then to apply special software to facilitate categorizing and coding the responses. There was not an exercise in "artificial intelligence"; the categorization and coding are entirely the results of judgements by the human coder. The computerization was designed simply to streamline the examination and processing of the textual responses.

2. Major findings and conclusions

2.1. Improved understanding of datasets

From the item response analyses, findings concern interpretation of item difficulty. When we construct educational achievement tests, there is always a conflict between wanting to provide great detail and richness through a finely articulated item set and wanting to get some kind of overall score or indication of general achievement. In a student testing program, the overall score may be very important—for example, for selection and placement. In a program assessment, the individual item statistics are suppose to be more important—for example to guide curriculum evaluation and development. In either case, but at different levels of generality, one is concerned with whether a set of items can be said to measure a single concept. We know that two items would measure the same concept only if they were identical, but the question is one of approximating uni-dimensionality. A major finding in the project analyses of item statistics graphics is that our classification of items is very shaky, even in very controlled studies like the Second IEA Mathematics Study, or the Ontario school board replications. Put four items together that seem to measure, for example, ability to do percent problems, and we find that (a) the difficulties vary, (b) the opportunities to learn vary, and (c) the sensitivity to opportunity to learn varies. A reasonable conclusion is that we need better theoretically-derived linkages between item categories if we are to connect item assessment data with prospects for teaching and learning.

From the school effects analysis, the findings are mainly descriptive. That is, there has been no real improvement in understanding why achievement is higher in one classroom or school than another. Perhaps such understanding is beyond the power of research using survey data. But from the point of view of description, the findings are interesting. First of all, it is clear that the school effects distribution is three-dimensional—we can differentiate classrooms according to inputs, outputs, and slopes (increments in output per unit input). That is, schools and classroom are seen to vary in what kind of student achievement they receive, in what kind of student achievement they produce, above and beyond what they receive, and in how much of an accelerating or flattening effect they apply to the student achievement distribution. This would seem to have important implications for educational policy. What is the policy on educational achievement distribution? It will have to account for all three dimensions.

A further finding from the school effects analysis is that within the three-dimensional configuration of school achievement, there are some easily identified special cases. In Ontario, these are the French schools and the private schools. The pattern of educational achievement and growth is dramatically different from those of other schools. This was not so obvious on the margin—that is, just looking at mean achievement. In the U.S.A., the "remedial" classes, which occur in grade 8, are a special category.

From the analyses of standardized tests, an important finding is that there is a small but critical proportion of bad data. This may derive from students who lose track of the answer sheet numbers or who cannot really read the test or understand the instructions. But the conclusion is that when standardized tests are employed, there must be very good statistical control to ensure that influential decisions about individual students are not made on garbage data.

From the the higher-level analyses and displays, especially for the Ontario school board replications of the IEA surveys, one finding is that there are few findings. That is, when we operate at a relatively high level of aggregation, such as over a comprehensive school board area, and when there is, in effect, a standard curriculum, then the differences in achievement and in opportunity are very small. This contrasts completely with the finding that schools and classrooms vary tremendously in achievement patterns. The conclusion for the policies of educational assessment may be that it is not worthwhile to carry out expensive testing in conditions that will hide rather than reveal variability.

2.2. New methodological insights

The topic of multilevel analysis is undergoing considerable development at this time. Some major statistical and computing breakthroughs have opened new possibilities for rigorous statistical estimation of school and classroom effects in educational survey data.

There are two findings from this project that provide some methodological cautions. First, the multilevel data can have outlier points that will destroy a non-robust analysis procedure. Second, within a seemingly homogeneous sample of multilevel data there can be lurking categories of points that correspond the qualitatively different educational circumstances and that produce dramatically different data structures. The conclusion is that robust analysis and careful graphics are essential, especially for the computer-intensive new statistical procedures.

From the item response analysis, there is a methodological insight into item parameters. It seems that a dimension of sensitivity to instruction may need to be conceptualized and brought in as an item response parameter.

The experience with coding open-ended responses may encourage the use of more flexible item formats in achievement and in attitude-opinion surveys.

2.3. Practical software

There was some expectation that this project would lead to specific new computer packages for analysis and display of assessment survey data. Except for the "S" macros for multilevel data and the special program for categorizing open-ended responses, that has not occurred. The reason for this is that the approach to graphics and data analysis has shifted.

For conventional statistical analysis, there is no reason to use anything beyond standard statistical packages, such as SPSS or SAS.

For the kinds of exploratory and experimental analyses that have been the object of this research project, the computing solution is seen now to be eclectic. That is, there are a number of superb statistical and graphical tools, such as MacSpin on the Macintosh and "S" on the SUN workstation. The goal of the experimental analyst is to learn to use these tools and to link them together. As new techniques are discovered and investigated, they will be incorporated into the standard packages.

2.4. Continuing work

No further funds were requested for this project. The equipment is in place and being used for continuing research in the same and in different domains (esp., hypertextual analysis of computer-mediated communication in education). It is partly incorporated into a new Advanced Computing Laboratory project and is available to other faculty and to advanced students through the OISE Computing Services Group.

The facility, together with some supplementary gear, will be used during 1988-89 in a major data banking and analysis project for the IEA Second Science Study, under a grant from the U.S. National Science Foundation and in cooperation with the University of Hamburg and Columbia University.